Electronic linkage and interrogation of administrative health, social care, and criminal justice datasets

Higgins, Cassie; Matthews, Keith

Published in:
Informatics for Health and Social Care

DOI:
10.1080/17538157.2020.1793346

Publication date:
2020

Document Version
Peer reviewed version

Electronic linkage and interrogation of administrative health, social care and criminal justice datasets: feasibility concerning process and content

RUNNING HEAD: Electronic linkage of health and social care data

Cassie Higgins and Keith Matthews

Division of Molecular and Clinical Medicine, University of Dundee, Mailbox 6, Level 6, Laboratories Block, Ninewells Hospital and Medical School, Dundee, DD1 9SY. Tel: +44 (0)1382 383025.

Corresponding author: Cassie Higgins

ORCID details and email addresses

Cassie Higgins, PhD. ORCID: 0000-0002-5506-324X Email: c.z.higgins@dundee.ac.uk

Keith Matthews, PhD. ORCID: 0000-0002-4478-5888. Email: k.matthews@dundee.ac.uk

Word count: 7917

This is an Accepted Manuscript of an article published by Taylor & Francis in Informatics for *Health and Social Care* on 24 July 2020, available online: https://doi.org/10.1080/17538157.2020.1793346.

Electronic linkage and interrogation of administrative health, social care and criminal justice datasets: feasibility concerning process and content

Abstract

The objective was to test the feasibility of a novel model of electronic linkage and interrogation of large, sensitive, administrative datasets derived from healthcare, social care and criminal justice. Participants comprised all individuals having completed suicide or drug-related death in Tayside between 2009 and 2014. Data were hosted, linked and pseudo-anonymized by a Trusted Third Party and were interrogated via secure access to the HIC Scottish Government-certified Safe Haven. Several barriers were encountered concerning data access, with all but one issue (obtaining criminal justice data) ultimately soluble. However, each barrier led to a substantial delay in either obtaining the required approvals or in receiving the specified data extracts. Generally, data coverage was good but data quality was poor, with almost a fifth of the data fields (17%) being less than 10% complete. Feasibility of this novel approach was demonstrated. Critically, this was achieved because of the central involvement of a Trusted Third Party and the use of a Government-certified Safe Haven. Future studies using a similar model of data acquisition and analysis should consider the potential delays resulting from organizations' lack of familiarity with their data-sharing protocols and procedures.

Keywords: health informatics; electronic data linkage; safe haven; data governance.

Introduction

Context

Some important clinical outcomes, for example, suicides and “drug deaths”, cannot be fully explained or predicted using only healthcare-derived data and, in consequence, there is an urgent need to test whether electronically linked data from partner agencies, for example, from social care and criminal justice systems, may improve our understanding and support meaningful prediction. There are recognized barriers to the electronic linkage of health and non-health datasets for such purposes. One of the key issues is the need to obtain approvals to hold and to interrogate identifiable data extracts, in order that relevant individuals in any study cohort can be identified across different administrative datasets. The key concern and risk is that of loss of confidentiality. There is currently no established precedent for the release of identifiable extracts of routinely-collected administrative health data in Scotland for the purposes of linkage with other public datasets; therefore, there is a need to test novel models of electronic linkage that can respond to the data governance requirements of partner agencies. Another significant issue is the use of non-unified, agency-specific individual identifiers. Whilst healthcare datasets in the UK contain the NHS Community Health Index (CHI) number which permits the linkage of all data to a single individual, other statutory non-health datasets contain, instead, agency-specific individual identifiers. This presents a challenge when attempting to identify individuals across both health and non-health datasets for the purpose of data linkage.

Mindful of these limitations and in the context of the policy imperative towards integration of health and social care services in Scotland, both for commissioning and operational delivery, there is an urgent need to develop robust methods to develop capacity to exploit routinely-collected regional and national datasets. Whilst routinely-collected, administrative, clinical datasets tend to be of lower quality and completeness than data collected specifically for research purposes, the use of large regional or national datasets has the advantage of being

highly generalisable. Interrogation of these large, inter-agency, highly generalisable datasets could lead to the development of more robust predictive models of disease risk. The novel aspect of the present study was to examine the barriers and potential solutions to setting up this type of linked data repository where no specific provision currently exists.

Objectives

The core objective of the present study was to test the feasibility of a novel model of electronic linkage and interrogation of large, pseudo-anonymized, sensitive datasets – from the related agencies responsible for healthcare, social care and criminal justice services – with the eventual aim of identifying 12-month risk factors for death by suicide.

The core elements of the study design were to bring together available data for those individuals who had died between 01.01.2009 and 31.12.2014 and where the official recorded cause of death matched specific inclusion criteria as per below. Further, the deceased were required to be resident in Tayside area at the time of death. Health Informatics Centre (HIC) Services were asked to generate "*controls*" to match the key characteristics of the deceased "*cases*". This required that HIC Services use their population level data to identify individual Tayside residents who matched the deceased with respect to gender, age and estimated socioeconomic status [as indicated by Scottish Index of Multiple Deprivation (SIMD); https://www2.gov.scot/Resource/0050/00504809.pdf]. For each "*case*", four "*controls*" were generated. The controls were required to be alive at the date of death of their matched cases.

In order to be able to address specific questions relevant to suicide deaths and drug deaths separately, three target cohorts were defined. The first was "*probable suicide"*, using ICD-10 codes proposed and used routinely by Scottish Government Information and Statistics

Division (ISD). The second cohort was "*probable drug deaths"*, using ICD-10 codes based on the Scottish Government and ISD's "baseline" definition of drug-related death. The relevant ICD-10 codes for each of these target cohorts are described in Appendix I. The third cohort was an additional set of ICD-10 codes, recommended by the Manitoba Centre for Health Policy (MCHP, 2014); the aim was to use this cohort both as a comparator cohort to the first cohort (in addition to a matched controls comparator cohort) and to amalgamate *with* the first cohort to provide a more inclusive definition of "*probable suicide"*.

In the following sections we will describe the process of undertaking this electronic linkage study and address the barriers – and potential solutions – to achieving linkage and interrogation of these datasets.

Methods

Legislation governing the use of administrative datasets

Routinely collected health and social care data are protected by legislation that falls into two categories: primary legislation (Acts of the Scottish Parliament); and secondary legislation (detailed regulations issued by means of Scottish Statutory Instruments as directed by primary legislation). The General Data Protection Regulation (GDPR) and Data Protection Act 2018 are the current laws ensuring data protection in the UK. Furthermore, the Human Rights Act 1998 incorporates into UK law rights and freedoms guaranteed by the European Convention on Human Rights. This legislation has an impact on which records can be created, retained and accessed, and it is relevant to those wishing to access identifiable information. However, since the present study used only pseudo-anonymized data, current legislation did not restrict planned use of these data.

Data protection: identification of a Trusted Third Party responsible for hosting and pseudo-anonymizing data

The Health Informatics Centre (HIC; https://www.dundee.ac.uk/hic/hicservices/), University of Dundee, was the facility that was commissioned to host all data required for the present study. HIC provides one of the available Scottish Government-certified Safe Havens and, as such, is acknowledged as a Trusted Third Party (TTP) in health informatics. HIC is governed by rigorous Standard Operating Procedures (HIC SOPs; https://www.dundee.ac.uk/hic/datasecurityconfidentiality/standardoperatingprocedures/) and is subject to annual independent audits (https://www.dundee.ac.uk/hic/datasecurityconfidentiality/). HIC holds generic ethical approval, which covers all work involving their data, awarded by the East of Scotland Research Ethics Committee (EoSREC; https://www.nhstayside.scot.nhs.uk/YourHealthBoard/TheBoardanditsCommittees/EastofScotlandResearchEthicsService/index.htm). This approval is subject to annual independent review and is overseen by the HIC Governance Committee.

HIC holds a comprehensive collection of electronic health registers on all individuals in Tayside who are registered with a primary care general medical practice (>99% of the population). HIC receives regular regional data extracts from national health registries held by Information Services Division (ISD), NHS Scotland, providing a complete record of all healthcare contacts for all patients. Data are available for the past 9-25 years, depending on data source, and span such areas as: community-dispensed prescribing; inpatient and outpatient treatment episodes and laboratory results. Data from different NHS healthcare services can be linked electronically since all NHS datasets are indexed by the Community Health Index (CHI) number, a unique 10-digit numerical patient identifier. Prior to release to

researchers, all data are pseudo-anonymized – CHI numbers are replaced with proxy CHI (proCHI) numbers, which are arbitrary alpha-numeric strings; meaningless outside of the HIC environment. Whilst researchers access data that has been fully anonymized, the overall process involved, whereby HIC holds both the data and the index information, results in pseudo-anonymization.

Data are accessed exclusively by approved and information security-trained researchers via a remote virtual desktop (Citrix XenDesktop, Santa Clara, CA). Interrogation of datasets is undertaken entirely within the HIC Safe Haven using pre-installed software. Analysis results can be exported following review by dedicated HIC personnel; however, no information can be copied directly from the virtual desktop to local desktops or portable storage devices.

Data sources

In addition to utilizing HIC-hosted datasets, the present study aimed to acquire regional data extracts from nationally and locally-held, sensitive datasets that were each subject to bespoke governance standards. Data were sought to cover the 12-month period preceding the index date for every individual included in the study. All data sources used in the present study are described below.

Scottish Suicide Information database (ScotSID)

ScotSID is a register held by ISD that acts as a central repository for information (from a variety of sources) on all probable suicides in Scotland. ScotSID was initiated in 2009 and provides information such as date of death, cause of death, demographic information and previous healthcare contacts. Further information can be found at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=101.

National Drug-Related Deaths Database (NDRDD)

NDRDD is a register held by ISD that acts as a central repository for information (from a variety of sources) on all drug-related deaths in Scotland. NDRDD was initiated in 2009 and provides information such as details of death, demographics, known substance misuse, previous overdose and previous healthcare contacts. Further information can be found at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=26.

Tayside Drug-Related Deaths Database (TDRDD)

TDRDD is governed by NHS Tayside's Public Health Directorate which acts on advice from the Tayside Drug Death Review Group (TDDRG), a collaboration of professionals representing the areas of health care, social care, criminal justice, the Third Sector and the three Tayside Alcohol and Drug Partnerships (ADPs). TDRDD is similar to NDRDD in terms of content and, therefore, contains data items concerning the cause of death and the personal and clinical characteristics of all cases included on the register.

National Records of Scotland: death certification (NRS Death)

This dataset is governed by the National Records of Scotland (NRS), a department of Scottish Government, and it contains data on all persons whose death was registered in Scotland. For each entry on this register the date and cause of death are recorded. Cause of death is recorded as one or more ICD code(s). Further information can be found at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=13.

Scottish Morbidity Registers: inpatient treatment (SMR01 and SMR04)

The Scottish Morbidity Registers (SMRs) are managed by Information Services Division (ISD) and governed by National Services Scotland (NSS), NHS Scotland. SMR01 provides a complete record of all acute general inpatient events and SMR04 provides a complete record of all psychiatric inpatient events. For example, both registers include data concerning admission reason and status, length of stay, treatment administered and disposal (including internal transfers to different specialties). Further information can be found at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=5 for the SMR01 and at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=7 for the SMR04.

Scottish Morbidity Register: outpatient treatment (SMR00)

SMR00 contains a record of all NHS outpatient clinic appointments: specialty of attendance; date of appointment; and attendance category of patient (which enables the calculation of "did not attend" (DNA) rates). Further information can be found at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=4.

Scottish Morbidity Register: cancer treatment (SMR06)

SMR06 provides a register of all patients in Scotland diagnosed with malignant disease. This register includes information on the clinical status of the disease, treatment administered and whether or not patients have received a terminal diagnosis. Further information can be found at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=8.

Scottish Morbidity Register: substance misuse (SMR25)

SMR25, also known as the Scottish Drug Misuse Database (SDMD), provides a register of all patients in Scotland who are in treatment for drug dependence or drug abuse. The majority (>90%) are in receipt of opioid agonist therapy (OAT) for the treatment of opioid

dependence; however, many have problems with multiple substance misuse. The register contains information concerning personal and domestic circumstances (including employment status and living circumstances), substance misuse (name of drug, amount consumed, frequency of consumption and route of administration) and medical treatments administered. Further information can be found at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=1.

Scottish Morbidity Register: maternity (SMR02)

An SMR02 record is opened every time a woman receives inpatient treatment for an obstetrics event during the perinatal period. This dataset comprises a range of possible data: diagnostic information; previous pregnancies; fetal terminations; proposals for delivery procedure; record of labor; baby record; and known drug and alcohol misuse. Further information can be found at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=6.

Scottish Birth Record (SBR)

SBR, formerly SMR11, is managed by ISD and governed by NSS, and it provides a record of all live and still births in addition to antenatal and post-birth events. SBR covers the first year post-birth, and individual case records contain up to 400 data items including gestation, birth weight and congenital abnormalities. Further information can be found at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=2.

NHS24 dataset

This dataset is managed by NHS24 and contains a record of all patient contacts with the service. Data items include: nature of call; advised course of action; ambulance dispatched; police dispatched; and involvement of psychiatric services.

Scottish Ambulance Service (SAS) dataset

This dataset is managed by SAS and provides a record of all patient contacts with the service. SAS records several data items: scores on the Glasgow Coma Scale; if patients are under the influence of drugs and/or alcohol; known history of substance misuse; police attendance required; known prescription medication; and administration of antagonist medication to treat overdose (naloxone to treat an opioid overdose or flumazenil to treat a benzodiazepine overdose).

Accident and Emergency (A&E) Datamart

A&E data are managed by ISD and governed by NSS, and the data items include information on presenting complaint, clinical status on presentation (including evidence of drug or alcohol misuse), treatments administered and disposal (including outpatient referrals). Further information can be found at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=3.

Prescribing Information System (PIS): community-dispensed prescribing

PIS is managed by ISD and governed by NSS, and it contains a record of all community-dispensed prescriptions. Data include: prescriber and dispenser details; name, strength and formulation of medication; directions for use; and British National Formulary (BNF) classification codes describing intent to treat. Further information can be found at https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=9.

NHS Tayside laboratories dataset

These data are owned by NHS Tayside and data governance lies with NHS Tayside's Caldicott Guardian. Collectively, this dataset comprises information from several laboratories: biochemistry; hematology; immunology; microbiology; and virology. Composite extracts provided information concerning the date, nature and result of each test.

Project-specific datasets held within the HIC environment

The Vascular Laboratory dataset provides the results of vascular CARSCAN and SEGPRES for all patients in Tayside dating back to 2000 and the results of vascular Duplex Doppler for all patients in Tayside dating back to 2008. The ECHO cardiogram dataset was incepted in 1994 and contains the results of all echocardiographic examinations for patients in Tayside. The Renal Register contains a record of all dialysis and transplant patients in Tayside dating back to 2002. The Scottish Care Information - Diabetes Collaboration (SCI-Diabetes) dataset contains a record of all patients with diabetes who were recruited to this study. Data are available dating back to 1996. The Tayside Allergy & Respiratory Disease Information System (TARDIS) contains a record of all patients with either chronic obstructive pulmonary disease (COPD) or lung cancer who were recruited to this study. Data are available dating back to 2001. Further information on these datasets can be found at https://www.dundee.ac.uk/hic/datalinkageservice/datasetinventory/#!faq-0.

Local Authority datasets: Social Work Departments

Local government in Scotland is comprised of 32 local authorities, designated as "councils" in accordance with the Local Government (Scotland) Act 1994. The three Social Work Department datasets are governed by the three respective Local Authorities within NHS

Tayside: Angus Council; Dundee City Council; and Perth & Kinross Council. All clients undergo at least one Needs Assessment; details of these interviews are recorded along with referrals to internal teams and external agencies. Internal teams maintain an ongoing record of assessments, interventions and outcomes for all clients referred to them following the initial Needs Assessment. Internal teams focus on the areas of: community care; child protection; and criminal justice.

Local Authority datasets: Finance Departments

These three datasets are managed by the three respective Local Authorities within NHS Tayside; however, some of the data items are governed jointly with the Department of Work and Pensions (DWP). The data contained within these datasets focuses on benefit entitlement and level of benefit awarded.

Data held by Police Scotland

These data are governed by Police Scotland and cover an extensive range of data items spanning criminal activity (relating to perpetrators, victims and witnesses), known associations, known health and substance misuse issues, detention pending court appearance and trial outcome. Further information can be found at http://www.sipr.ac.uk/.

Results

This section begins by examining "process" issues (i.e. obtaining approvals and data extracts) and concludes with an examination of "content" issues (i.e. data quality and data coverage).

Process: obtaining required approvals and receipt of data extracts

Due to the extensive range, obtaining data extracts was associated with a complex set of required approvals. Prior to study initiation, approval was sought and obtained from NHS Tayside's Research & Development Department. An application for ethical approval was not required since HIC holds generic ethical approval for all work undertaken within their service; however, a Favorable Ethical Opinion was sought and obtained from the East of Scotland Research Ethics Committee (EoSREC). The requirement for additional approvals is shown in ***Table 1***, and the nature of these approvals is described further, below.

[Insert Table 1 around here]

PBPP approval

Datasets that are considered to contain sensitive information required approval from the Public Benefit and Privacy Panel for Health and Social Care (PBPP). PBPP is the governance structure of NHS Scotland that has responsibility for governance-related matters and acts on behalf of NHSScotland Chief Executive Officers and the Registrar General.

The key issue that arose from the application was concern around the inclusion of personal identifiers in these sensitive datasets, data that would then be used to identify the relevant extracts from all other datasets. Concern fell around the proposed method of making case identifiers known to other host agencies. This was a requirement of the present study, since the aim was to link data for individuals across multiple datasets. A resolution was achieved that satisfied all partner agencies – extracts were obtained from all agencies for all events during the specified timeframe, and HIC undertook the final extraction (of appropriate cases and controls) within the HIC environment. This ensured that no other partner agency became

aware of the “probable suicide” and “drug-related death” classifications of individuals’ deaths.

Whilst there were no other particular issues associated with the application, the process took considerably longer than was anticipated. Initial contact was made on 27/10/14 and formal approval was received on 25/04/16. HIC received the ScotSID extract on 12/09/16 and data were made available to the research team on 07/10/16. HIC received the NDRDD extract on 15/12/16 and data were made available to the research team on 22/12/16. Subsequent to submitting the application, the research team made the decision to withdraw SAS from the battery of requested datasets. This decision was based on poor data quality, and it is described later in this section. In June 2015 the PBPP disclosed that they would be unable to provide an extract of the NHS24 dataset for the required period of time and advised that the only recourse would be to arrange to receive a direct data feed from NHS24. An overview of the time scale relating to PBPP approval, and subsequent data acquisition, is shown in ***Figure 1***.

[Insert Figure 1 around here]

Information Sharing Protocols (ISPs)

Initial contact was made with NHS24 in June 2015, and the NHS24 Executive Committee formally approved the data request on 27/09/16; however, the data extract was not received prior to the conclusion of the project.

The Scottish Ambulance Service (SAS) stipulated that, in addition to PBPP approval, an individual ISP would also be required. Following discussions with SAS, it was decided that

this dataset request should be removed from the PBPP application. This decision was based on poor data quality. SAS reported that only 37.7% of their patients were associated with an NHS CHI number during the observation period (i.e. were individuals that could be identified). Following further discussion, it transpired that those who were unconscious or incoherent were least likely to be identifiable in the SAS dataset. Since data concerning these individuals could not be considered to be missing at random (MaR), the SAS dataset was removed from the required list of extracts.

The Tayside Outpatient Appointments System (TOPAS) is managed by a non-proprietary agency (Cambric); therefore, an ISP was negotiated with Cambric acting on behalf of NHS Tayside. The appropriate protocols and procedures were identified rapidly; however, the non-proprietary administrative costs rendered this extract cost-prohibitive.

The Scottish Institute for Policing Research (SIPR) was engaged at the outset of the project to ensure familiarization with governance protocols and procedures. SIPR was established in 2007 following investment from both the Scottish Funding Council and the Association of Chief Police Officers in Scotland. It represents collaboration between Police Scotland and 14 Scottish universities and takes responsibility for initial review and consultation regarding proposed research studies involving data held by Police Scotland. Following approval from SIPR, initial contact was made with Police Scotland on 01/09/14 and interim formal approval was awarded on 16/10/15 by Strategic Planning Development, pending the establishment of a satisfactory Police Scotland Minute of Agreement (MOA). Police Scotland presented the proposed MOA to SIPR for review and consultation. In the meantime, in order to avoid further delays, Police Scotland examined the terms of the project's EoSREC approval to facilitate an agreed Standard of Behavior relating to information handling and data security

(i.e. the pragmatics associated with data transfer). Following internal discussions, however, between SIPR and Police Scotland, concerning the MOA, formal approval was repealed in November 2016. The project was informed that the proposed pseudo-anonymization by the TTP was unsatisfactory and, in order to proceed, a method of true anonymization was required. In the final month of the project, the research team began working with Police Scotland on a new method of information handling; however, it was clear that this would not be achieved prior to the conclusion of the study, and that the identification of a satisfactory method could only be applied in future studies.

Obtaining an extract of the Tayside Drug-Related Death Database (TDRDD) was problematic, largely due to a lack of established access protocols and procedures. Whilst NHS Tayside's Public Health Directorate acts as the gatekeeper for this dataset, they stipulated that Tayside Drug Death Working Group (TDDWG) approval was required in order to proceed with the application. TDDWG involvement resulted in lengthy delays in obtaining access to the required data. One further issue resulted from the decision that presence on the TDRDD could be indicated; however, no other data fields would be made available to the project. This meant that individuals from the TDRDD had no additional data from that dataset (e.g. toxicology findings, previous overdoses, access to take home naloxone (THN) socioeconomic information and personal characteristics, etc.) and, therefore had to be excluded from some analyses.

Initial contact was made with the three Local Authority Social Work Departments in August 2014. Angus Council Social Work Department (ACSWD) seemed unclear about protocols relating to data sharing and, in consequence, meetings were required with numerous staff until there were no avenues remaining open. At that point, the research team called upon the

assistance of a former colleague who was able to help in facilitating contact with senior staff within ACSWD. The ISP was signed on 9 December 2016; however, no data extract was received prior to the conclusion of the study. Contacts within Dundee City Council Social Work Department (DCCSWD) were aware of protocols and procedures for sharing their data. Formal sign-off was obtained in August 2015 and the DCCSWD indicated that the transfer could be made at any point thereafter, on our request. This data extract was dependent upon having first identified the cohort; therefore, due to delays in obtaining the core datasets, this extract from DCCSWD was finally transferred in January 2017. Perth & Kinross Council Social Work Department (PKCSWD) seemed unclear about protocols relating to data sharing; however, staff expressed engagement with the aims of the study and a desire to contribute. A responsible officer was identified on 26/11/15; however, this individual did not respond to any further correspondence. Instead the officer sent approval, prompted by senior colleagues, to a colleague outside of the research team who informed us of this in December 2017. The extract was transferred to the TTP in January 2017.

Initial contact was made with the three Local Authority Council Tax & Benefits Departments (LACTBs) in December 2015. Representatives from the three LACTBs reached the decision to work together with the research team. This approach seemed to have eased the process and encouraged LACTB engagement with the aims of the project. There were, however, some issues concerning access to certain data variables. This was a result of joint gatekeeper responsibilities between the LACTBs and the Department of Work and Pensions (DWP). Having no experience of the index system that we proposed to use in incorporating non-CHI linked data, the LACTBs were required to undertake internal consultations prior to agreeing an ISP. No data extracts were received prior to the conclusion of the study.

Initial contact was made with the Department of Work and Pensions (DWP) in August 2014. Several attempts were made to engage the DWP with the core aims of the study; however, this was not achieved. On each occasion, a message was received informing the study team that the DWP would identify an appropriate member of staff, and that that person would make contact with a member of the study team.

During the initial months following study inception, discussions were undertaken regarding the potential to obtain diagnostic information from Primary Care datasets. We were informed that the data may not be reliable; however, the greatest issue was that each GP practice acts as its own gatekeeper. This meant the need to construct ISPs with each of the 62 individual practices in Tayside, assuming consent to participate was given. Whilst seeking this level of consent would be resource-intensive, the greater problem was the projected consent figures. This led to concerns in the project team that the inclusion of (consented only) GP practice data could skew the findings. In light of the required resources and the potential to skew finding, the decision was taken not to include Primary Care data in the present study.

NHS Tayside Caldicott Guardian approvals

An NHS Caldicott Guardian is a senior officer who holds responsibility for protecting the confidentiality of individuals' health data. Each health board within Scotland has a Caldicott Guardian who is responsible for data collected within that health board area, and there is an NHS Scotland Caldicott guardian who is responsible for national data. The present study utilized data from NHS Tayside and, in consequence, sought NHS Tayside Caldicott approval. NHS Caldicott Guardians are governed by the seven Caldicott Principles, which apply to the handling of patient-identifiable information (https://www.igt.hscic.gov.uk/Caldicott2Principles.aspx).

Obtaining Caldicott approval was relatively straightforward, and it was obtained for: the HIC-hosted extracts; TOPAS; SMR25; and TDRDD. The HIC-hosted datasets were made available as required. As discussed previously, TOPAS non-proprietary administrative costs rendered this extract cost-prohibitive. Caldicott approval for TDRDD was non-problematic, but the requirement for an independent ISP led to substantial delays, as discussed previously. Finally, formal Caldicott approval was obtained for the Scottish Morbidity Register on substance use disorders (SMR25); however, due to upload issues between the Tayside Substance Misuse Service and ISD, this extract was never delivered to the TTP.

Gatekeeper approvals

The project-specific datasets held by HIC Services required approval from individual data gatekeepers – usually the study's Principal Investigator (PI). In each case this was achieved by email correspondence, which was adequate approval to facilitate the release of these datasets, within the Safe Haven environment, to the present study team.

Process: data integration within the TTP Safe Haven

An overview of the flow of data into the TTP Safe Haven is shown in ***Figure 2***.

[Insert figure 2 around here]

Obtaining extracts of the core datasets

The core datasets (i.e. those used to identify cases in the study) were uploaded to HIC Services using the host agency's preferred means of secure data transfer. No problems were encountered during this part of the process.

Obtaining NHS dataset extracts

An extract (based on the dates of the observation period) from each of the NHS datasets (i.e. those containing CHI number identifiers) was uploaded to HIC Services. These extracts contained data for all individuals from within the specified timeframe. Within the HIC environment, data were extracted for all cases (based on the identifiers contained within the core datasets) and controls (selected from the general population by HIC Services), and the remainder of each dataset was destroyed.

Obtaining data from non-CHI-indexed sources

Prior to the inflow of data from non-CHI-indexed sources, a novel indexing procedure was used to facilitate the secure transfer of a minimum dataset in each case. The testing phase was successful in all cases. With Police Scotland having withdrawn from the study, this was applied to the three Local Authority Social Work Departments and the three Local Authority Council Tax & Benefits Departments. The indexing procedure is shown in ***Figure 3***.

[Insert figure 3 around here]

Figure 3 shows that non-CHI-indexed datasets, were indexed (with the index key of CHI numbers and pseudo-identifiers held by HIC Services). In practice, this meant that each external dataset contained a new variable: the pseudo-identifier, an arbitrary alpha-numeric string which was meaningless outside of the HIC environment. Within HIC Services, the case and control CHI numbers were then compared with the key and relevant individuals were selected (using the pseudo-identifiers, rather than the host agency identification number). A hard copy of pseudo-identifiers was taken to each host agency by a member of

HIC personnel, who then extracted the relevant data from each of the host agency dataset and transferred the extract securely to HIC Services, without leaving a footprint on the host dataset of records accessed. None of the extracts from non-CHI-indexed datasets was received prior to the conclusion of the study; however, two extracts were received during the following months. No problems were encountered and feasibility was demonstrated.

Process: electronic linkage within the HIC environment

As described above, the 'core datasets' were those that were used to identify cases in the study (i.e. individuals that completed suicide during the observation period), and matched controls were identified by HIC using their 'CHI Register', which also contains demographic information for all individuals with an NHS CHI number, thus facilitating the matching process. Since regional extracts of all of the required NHS datasets are routinely hosted by HIC and updated on a regular basis, the NHS CHI number was used to extract information from these datasets for all relevant cases and controls within the HIC environment. Thereafter, data were pseudo-anonymized prior to release to the research team. This process involved constructing a key which enabled the replacement of the NHS CHI number with a random alpha-numeric string, meaningless outside of the HIC environment. The data were pseudo-anonymized using this key, rather than fully-anonymized, to facilitate further linkage by the research team within the Safe Haven, as described in the following section. As described previously, the (non-CHI-indexed) Local Authority data extracts were transferred to HIC in a pre-pseudo-anonymized form. Linkage was relatively unproblematic and it was found that all pseudo-identifiers were assigned appropriately. The principal issue encountered at this stage was that, in some, but not all cases, data had been extracted by HIC Services for dates that fell beyond those specified. For example, diagnoses of diabetes mellitus spanned each individual's lifetime; whereas, diagnoses of malignant disease spanned

only the observation period. In some cases, therefore, inappropriate data were deleted from datasets.

Process: data preparation and electronic linkage within the Safe Haven

All data extracts relevant to the presented study were uploaded, within the HIC environment, to the HIC Safe Haven, where all data preparation and interrogation took place. Datasets were presented in the form of comma separated values (CSV) files, ready to be imported to any statistical software. Data were prepared and analyzed by the research team within the HIC Safe Haven virtual environment. The Safe Haven is accessed via a secure web link (which requires a Citrix plugin for activation), and a password-protected login is required to access study data.

First, each dataset was cleaned and coded. Very few of the 830 data variables in the datasets were coded appropriately for statistical analysis and every variable had cells containing null values. In consequence of this, data cleaning and coding took substantially longer than anticipated.

Second, data were transformed. This largely involved the transformation of datasets from long- to wide-form. In long format each line in the dataset represents an event; however, in wide format each line represents an individual. This process required additional coding; however, no problems were encountered during the data transformations.

Third, data were linked across datasets. This was achieved as a result of the process of pseudo-anonymization, since it facilitate the linkage of events for each participant across multiple datasets. In order to respond to specific hypotheses, data were extracted from

multiple datasets and linked to form one relevant dataset for each research question. All variables remained intact and no new variables were constructed by triangulation of multiple variables. This was because data collection procedures could not be confirmed as being identical.

Content: data quality and data coverage

The quality of data varied both within and between datasets. Of the 830 data fields, none were fully-completed, including those identified as "mandatory" by host agencies. Data field completion rates were higher for social care than for health datasets, and these datasets were more likely to be coded appropriately for the purpose of statistical analysis. An overview of the proportion of completed fields is shown in ***Table 2***.

[Insert Table 2 around here]

Table 2 shows that almost a fifth of the data fields (17%) were less than 10% complete. This was not a result of any aspect of the linkage processes since all data fields remained intact throughout the study. It should be noted that most of these data fields contained no data. An additional problem was that, where null values were returned in cells, it was not always clear if this represented a negative response or was indeed truly "missing" data. Furthermore, it was not clear where missing data were likely to be an artefact of administrative systems/lack of time/etc. and where they may have reflected under-represented subgroups and could skew findings.

Data coverage was generally satisfactory. From the perspective of undertaking health research; however, three key data variable were obvious in their absence. First, diagnostic

information was not contained within routinely-collected, nationally-held datasets. Indeed, this information is not stored for psychiatric morbidity, at least, in a consistent electronic format. Secondly, as discussed previously, impracticalities meant that Primary Care data were not included in the present study. Finally the DWP was engaged as a partner agency with the aim of ascertaining financial hardship through receipt of a qualifying benefit; however, as described previously, these data were not forthcoming.

Overview of the integrated datasets held within the HIC Safe Haven

The complete battery of datasets totaled 27 and included 1528 cases and 6112 controls. The number of events totaled over quarter of a million; however, this figure was largely influenced by dispensed prescription drugs and, to a lesser degree, laboratory results.

Discussion

Barriers to approvals and receipt of data extracts

The first barrier to data access was encountered during the application stages, whereby many agencies were unclear about the existence of established precedents and the relevant personnel within their agency for handling data-sharing requests. This resulted in substantial delays. Numerous strategies were employed in assisting partner agencies through the various stages of the application process, based on the previous experience of the research team. This included, but was not limited to, agency-wide presentations by the research team, question and answer sessions with partner agencies, meeting with several personnel within agencies in efforts to ensure that the most appropriate members of staff were identified, immediate responses to all queries from partner agencies to ensure continued momentum, sending reminders of action required by partner agencies, and so on. One of the most effective

strategies employed was to use a “top-down” approach when communicating with external agencies – i.e. to begin as high up the chain of command as possible.

A significant barrier arose when the PBPP communicated that it would be unwilling to disclose individual patient identifiers for the purpose of obtaining data extracts from other partner agencies covering the relevant individuals. The resolution to this barrier involved further negotiations with the PBPP and the use of a TTP, in this case HIC Services. The data were transferred to the TTP with personal identifiers intact and all data extractions were undertaken by the TTP, either within their virtual environment or using the HIC-generated pseudo-anonymization index key. The PBPP was satisfied with this proposal and no further problems were encountered around this issue. Just prior to the initiation of the present study, the PBPP had replaced the former Privacy Advisory Committee (PAC) and this may have contributed to the substantial delay in processing our application; however, the new PBPP appeared to be relatively poorly informed concerning the data that could or could not be made available through its governance system. Several months after having included NHS24 in the PBPP application, the research team was informed that NHS24 data could not be made available through the PBPP, and that an independent ISP would be required for access to NHS24 data. This lack of knowledge led to substantial delays in final receipt of the NHS24 data extract.

A further barrier was the potential financial cost in obtaining data extracts, for example, where non-proprietary IT systems were in place. This was the case for the three local authorities; however, each subsumed the non-proprietary administrative costs associated with data acquisition. The Tayside Outpatient Appointments System (TOPAS) is also managed by a non-proprietary agency (Cambric), and an ISP was negotiated with Cambric acting on

behalf of NHS Tayside. The appropriate protocols and procedures were identified rapidly; however, the non-proprietary administrative costs rendered this extract cost-prohibitive. The involvement of third partner agency staff raised a new set of problems: namely that these members of staff were accountable within their own agency and not directly accountable to the target agency. The result was that, even after having received sign-off, third agency staff had the opportunity to block or to delay data transfer. In some cases this was due to lack of prioritization; however, in the case of one Local Authority, the data transfer was blocked on alleged *"ethical grounds"*. Through discussions, it transpired that the third agency was not fully aware of all safety protocols, and there were no further problems; however, this resulted in further delays in obtaining data extracts.

Receipt of formal approval did not, however, necessarily indicate that the required data extract would be forthcoming, even where there was no third agency involvement. On several occasions, data managers not familiar with the authority of their governance departments stalled on delivering extracts expecting us to begin a dialogue with them in order to obtain their approval, despite that not being required. This was thought likely to be a function of the novelty of this study and, as more studies are undertaken using health informatics approaches to link data from social care and criminal justice sectors, precedents will become more readily established and understood. In the meantime, however, any future studies should factor into protocols additional "buffer" time to counteract delays in their anticipated timeframes. From the perspective of partner agencies, continued governance involvement and accountability could assist in efficient extract delivery.

The key continuing barrier to data access concerns those held by Police Scotland. Following receipt of formal approval from Police Scotland, their governance officers decided to

withdraw approval due to concerns around the pseudo-anonymization procedure and, more specifically, that the index key would be held by an external agency (i.e. HIC Services). Police Scotland has a well-established precedent for data sharing in the form of anonymized data feeds; however, this novel approach presented a challenge. In conclusion, Police Scotland would require a guarantee of full anonymization prior to considering data sharing using this model. The research team has since entered into further discussions with Police Scotland and is now in a position to test further protocols and procedures which can guarantee anonymity for individuals.

Barriers to data integration within the TTP Safe Haven

As discussed previously, the potential route to a key barrier concerning data integration within the TP Safe Haven was that agencies holding non-CHI-indexed data were concerned about transferring a data feed containing all data for all individuals during the observation period, and the PBPP was concerned about not making individual identifiers known to these agencies. As discussed, this required the use of HIC-generated pseudo-anonymization index keys. No problems were encountered, however, and proof of concept was demonstrated. The only issue encountered was that extraneous data were made available in the Safe Haven (i.e. for individuals in the study, but for dates spanning far beyond those of the observation period). This issue was highlighted to HIC Services and remedied rapidly. No other issues were encountered regarding data integration within the TTP Safe Haven.

Barriers to data interrogation

Data quality and coverage were the two key barriers encountered in terms of data interrogation. The findings of this study show that 46% of all data fields were less than half-completed and that almost a fifth of all data fields contained no data. None of the data fields

identified as “mandatory” by the partner agencies were fully-completed; however, the completed proportion was higher for mandatory fields than for optional fields. Non-completion of data fields was a greater issue in the health, than in the social care, datasets. It is difficult to speculate on potential solutions to this problem without further knowledge of the specific barriers faced by each agency in completing all data fields, or at least all mandatory data fields. Health informatics-based research is viewed as a strategic priority and asset. In order to keep pace with developments, health and social care agencies should be required over time to assume accountability for maintaining data quality beyond current levels.

Data coverage was generally good, with only a few obvious omissions; however, these omissions took the form of complete datasets, rather than data fields within any one dataset. In terms of health informatics research, perhaps the greatest issue is the lack of routinely-collected electronic information on patient diagnostic status, particularly in relation to psychiatric morbidity. The other obvious omissions were indicators of financial hardship (due to lack of DWP engagement) and Primary Care contacts (due to each practice acting as its own data gatekeeper, and the consequent high investment of resources required to engage GP practices as partner agencies). The future of health informatics depends on stakeholder engagement, established data sharing precedents, and partner agency accountability for data quality.

Feasibility regarding the identification 12-month risk factors for suicide completion

The feasibility of obtaining many of the required datasets was demonstrated. Obtaining data from Local Authority Finance Departments and Police Scotland was not achieved; however, with additional time, it is likely that these datasets would have been obtained. The datasets

included in the present study provided a wealth of data, enabling the derivation of a number of statistical models identifying 12-month risk factors for suicide completion in both clinical and non-clinical populations. The principal challenge to the feasibility of using routinely-collected data held within health registers was the poor completion of some of the key data fields. This necessitated the use of imputational techniques, where feasible, and resulted in the exclusion of some data fields that would have been included otherwise. Data completion was less problematic in routinely-collected social care data provided by Local Authorities. The findings from the primary data analyses were presented to the Scottish Government, with the aim of informing suicide prevention policy and further developing clinical practice in this area, and work is currently underway on several manuscripts which will be submitted to peer-reviewed journals for publication.

Cost of conducting the present study

The total cost of the study was the sum of the costs associated with using a Trusted Third Party, one researcher's full-time salary for 24 months and obtaining and pseudo-anonymising the non-NHS data (because these datasets used different person identifiers). The TTP took responsibility for hosting and pseudo-anonymising the relevant datasets. The total cost of this service was circa £60,000 over the total study period. The researcher took responsibility for familiarising herself with the required legislative and procedural components associated with obtaining these data, negotiating with the relevant partner agencies the specific data items to be obtained from each dataset, guiding the TTP regarding the specifics of their input, constructing satisfactory Information Sharing Protocols in conjunction with Local Authority, Police Scotland and University solicitors, co-ordinating the testing and final procedure associated with indexing non-NHS data, cleaning and coding data, and linking and analysing datasets. The total cost of the researcher's salary plus University overhead costs was £108k.

The external agency that indexed the non-NHS data took responsibility for implementing their previously-tested procedure in the Local Authority context and was engaged in discussions with Police Scotland concerning the implementation of a similar procedure using their data. The total cost of this service was £12k. The partner agency costs associated with working on this study were subsumed by these partner agencies: the Information Services Division of National Services Scotland, the three Local Authorities and Police Scotland.

Conclusions

With the use of a Trusted Third Party, feasibility was demonstrated for this novel model of electronic linkage and interrogation of large, sensitive datasets – from the disciplines of health, social care and criminal justice – with the aim of identifying 12-month risk factors for suicide or drug-related death. Most of the barriers to data access, linkage and interrogation were resolved; however, they resulted in substantial delays in the study timeline, and this was particularly true during the approvals stage of the project. Further studies in this area should be aware of the potential for substantial delays and adjust study protocols accordingly. Data quality was generally poor, and many data fields held no data, even when the host agency had identified it as a "mandatory" data field. Driven by policy and culture towards the increasing integration of health and social care commissioning and service delivery, healthcare, social care and criminal justice services should consider developing established data-sharing protocols and procedures, and also clear accountability for data quality. Ultimately, the feasibility of the present study was demonstrated; however, additional time would have been required in order to obtain all of the desired datasets. Poor completion of some of the key data fields in the health datasets was the most significant challenge in deriving statistical models of 12-month risk factors associated with suicide completion, and this necessitated the

use of imputational techniques, where feasible. However, there was a wealth of data that were used successfully in deriving statistical models.

Acknowledgements

The work was supported by Scottish Government with an award to Professor Keith Matthews and Dr Brian Kidd. Additionally, we would like to acknowledge the work of Dr Mark McGilchrist, University of Dundee, in facilitating the inclusion of non-CHI datasets within the present study and also Neil Fraser in Chairing our Project Stakeholders' Reference Group.

Disclosure of interest

Cassie Higgins has no interests to disclose.

Keith Matthews has chaired advisory boards for studies of Deep Brain Stimulation for Obsessive-Compulsive Disorder sponsored by Medtronic. He has received educational grants from Cyberonics Inc. & Schering Plough, and has received research project funding from Merck Serono, Lundbeck, Reckitt Benckiser, St Jude Medical and Indivior. He has received travel and accommodation support from Medtronic and St Jude Medical to attend scientific meetings.

References

Manitoba Centre for Health Policy (MCHP), University of Manitoba. Concept: Suicide and Attempted Suicide (Intentional Self Inflicted Injury). Winnipeg, Canada: University of Manitoba; 2014 [accessed 2014 Feb 6]. http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1183

Appendix I: ICD codes used to identify target cohorts

Probable suicides were identified in Tayside during the period 2009-2014 Identification of these cases was based on the Scottish Government criteria, where cause of death is reported as one of the following codes from the International Classification of Diseases (ICD v10):

- X60-64 and Y87.0 (intentional self-harm)
- Y10-34 and Y87.2 (events of undetermined intent)

Drug-related deaths were identified in Tayside during the period 2009-2014. Identification of these cases was based on the Scottish Government's "baseline" definition of drug-related death, where cause of death is reported as one of the following codes from the ICD v10:

- **F11** Disorders related or resulting from abuse or misuse of opioids
- **F12** Disorders related or resulting from abuse or misuse of cannabis
- **F13** Disorders related or resulting from abuse of misuse of sedatives or hypnotics
- **F14** Disorders related or resulting from abuse or misuse of cocaine
- **F15** Disorders related or resulting from abuse or misuse of other stimulants
- **F16** Disorders related or resulting from abuse or misuse of hallucinogens
- **F19** Disorders related or resulting from abuse or misuse of other psychoactive substances
- **X40-X44**[1] Accidental poisoning
- **X60-X6**[1] Intentional self-poisoning by drugs, medicaments and biological substances
- **X85**[1] Assault by drugs, medicaments and biological substances
- **Y10-Y14**[1] Event of undetermined intent, poisoning

[1] In the presence of at least one of the following T-Codes:

- **T40** Poisoning by, adverse effect of and underdosing of narcotics and psychodysleptics [hallucinogens]
- **T41** Poisoning by, adverse effect of and underdosing of anesthetics and therapeutic gases
- **T42** Poisoning by, adverse effect of and underdosing of antiepileptic, sedative, hypnotic and antiparkinsonism drugs
- **T43** Poisoning by, adverse effect of and underdosing of psychotropic drugs, not elsewhere classified

In order to ensure that all appropriate cases were included in each of the target cohorts, we triangulated data from different sources to construct the cohorts.

(1) ***Cohort 1 – "probable suicide"****:* Identified through a combination of all cases in the ScotSID dataset, plus appropriate ICD 10 codes in the NRS death dataset.

(2) ***Cohort 2 – "probable drug death"****:* Identified through a combination of all cases in the NDRD dataset, plus all cases in the locally held Tayside DRD dataset, plus appropriate ICD 10 codes in the NRS Death dataset.

(3) ***Other deaths related to "high risk behaviors"****:* Identified through appropriate ICD 10 codes in all four datasets (ScotSID, NDRDD, Tayside DRD and GRO Death dataset).

Table 1: requirement for specific approvals in order to obtain the required data extracts

Required approvals	Datasets
Public Benefit and Privacy Panel for Health and Social Care (PBPP)	Scottish Suicide Information Database
	National Drug-Related Death Database
	Scottish Ambulance Service
	NHS24
Information Sharing Protocols (ISPs)	NHS24
	Scottish Ambulance Service
	Tayside Outpatient Appointments System
	Police Scotland
	Tayside Drug Related Death Database
	Local Authority Social Work Department data (Angus, Dundee and Perth & Kinross)
	Local Authority Finance Department data (Angus, Dundee and Perth & Kinross)
	Department of Work and Pensions
	Primary Care
NHS Tayside Caldicott Guardian	Tayside Drug Related Death Database
	Tayside Outpatient Administrative System (TOPAS)
	HIC-hosted Scottish Morbidity Registers (SMR00; SMR01; SMR02; SMR04; SMR06; and SMR25)
	HIC-hosted NHS Tayside laboratory data

	(biochemistry; virology; hematology; immunology; and microbiology)
	Other HIC-hosted datasets (SBR; NRS Death; A&E; CHI Database; and Demographics).
Gatekeeper approval	Vascular Laboratory
	ECHO cardiogram
	Renal Register
	SCI-Diabetes
	TARDIS

Table 2: Percentage of data fields completed (by percentage completion of each data field)

Percentage completion of each data field									
0-9%	10-19%	20-29%	30-39%	40-49%	50-59%	60-69%	70-79%	80-89%	90%+
17%	4%	7%	9%	9%	16%	15%	14%	9%	0%

Figure captions

Figure 1: Timeframe for formal approval and receipt of data extracts, and status of application at study conclusion

Figure 2: Flow of data through the study framework

Figure 3: Indexing procedure used in obtaining non-CHI-indexed data: transfer methods and security protocols